Overview of The Virtual Data Center Project and Software

Micah Altman, L. Andreev, M. Diggory,
G. King, E. Kolster, A. Sone, S. Verba
Harvard University
G-4 Littauer Center, North Yard
Cambridge, MA 02138
(617) 496-3847
Micah_Altman@Harvard.edu

Daniel L. Kiskis and M. Krot
University of Michigan
Media Union
2281 Bonisteel
Ann Arbor, MI 48109

ABSTRACT
In this paper, we present an overview of the *Virtual Data Center* (VDC) software, an open-source digital library system for the management and dissemination of distributed collections of quantitative data. (see <http://TheData.org>). The *VDC* functionality provides everything necessary to maintain and disseminate an individual collection of research studies, including facilities for the storage, archiving, cataloging, translation, and on-line analysis of a particular collection. Moreover, the system provides extensive support for distributed and federated collections including: location-independent naming of objects, distributed authentication and access control, federated metadata harvesting, remote repository caching, and distributed 'virtual' collections of remote objects.

Categories and Subject Descriptors
H.3.7. [**Information Systems**] -Information Storage and Retrieval - Digital Libraries (H.3.7); -- *DL System Architecture, Distributed Systems, Information Repositories, Federated Search, DL Impact*

General Terms
Management, Design, Standardization

Keywords
Numeric Data, Open-Source, Warehousing.

1. INTRODUCTION
Researchers in social sciences, and in academia in general, increasingly rely upon large quantities of numeric data. The analysis of such data appears in professional journals, in scholarly books, and increasingly often in popular media. For the scholar, the connection between research articles and data is natural. We analyze data and publish results. We read the results of others' analyses, learn from it, and move forward with our own research.

But these connections are sometimes difficult to make. Data supporting an article are often difficult to find and even more difficult to analyze. Archiving, disseminating and sharing data is crucial to research, but is often costly and difficult. [Sieber 1991] Thus, our ability to replicate the work of others and to build upon it is diminished. Researchers, university data centers, and students all face challenges when trying to find and use quantitative research data.

The *Virtual Data Center* (VDC) software is a comprehensive, open-source digital library system, designed to help curators and researchers face the challenges of sharing and disseminating research data in an increasingly distributed world. The VDC software provides a complete system for the management and dissemination of federated collections of quantitative data.

2. Features, Design, and Implementation
The VDC provides functionality for producers, curators and users of data. For producers, it offers naming, cataloging, storage, and dissemination of their data. For curators, it provides facilities to create virtual collections of data that bring together and organize datasets from multiple producers. For users, it enables on-line search, data conversion, and exploratory data analysis facilities.

More specifically, the system provides five areas of functionality:

(1) Study preparation. (unique naming, conversion tools for multiple data and documentation formats, tools for preparing catalog records for datasets);
(2) Study management. (file-system independent dataset and documentation storage, archival formatting, cataloging);
(3) Interoperability. (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 query protocol support);
(4) Dissemination. (extract generation, format conversion, subset generation, and exploratory data analysis).
(5) Distributed and federated operation. (location-independent naming, distributed virtual collections, federated metadata harvesting, repository exchange and caching, and federated authentication and authorization)

Each VDC library comprises a set of independent, interoperating components. These independent VDC libraries can also be federated together, sharing collections through harvesting and dynamic caching processes. (See Figure 1)

Many of the core components are modeled after the design described in Arms [1997]: objects are stored in repositories, referenced through names that are resolved by name servers, and are described in index servers that support searching. We also incorporate some features of the NCSTRL [Davis 2000] architecture and particularly Lagoze's [1998] idea of a collection as an object that groups queries against remote index servers.

We have extended this core design in several significant ways. First, to support completely distributed operation of the

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

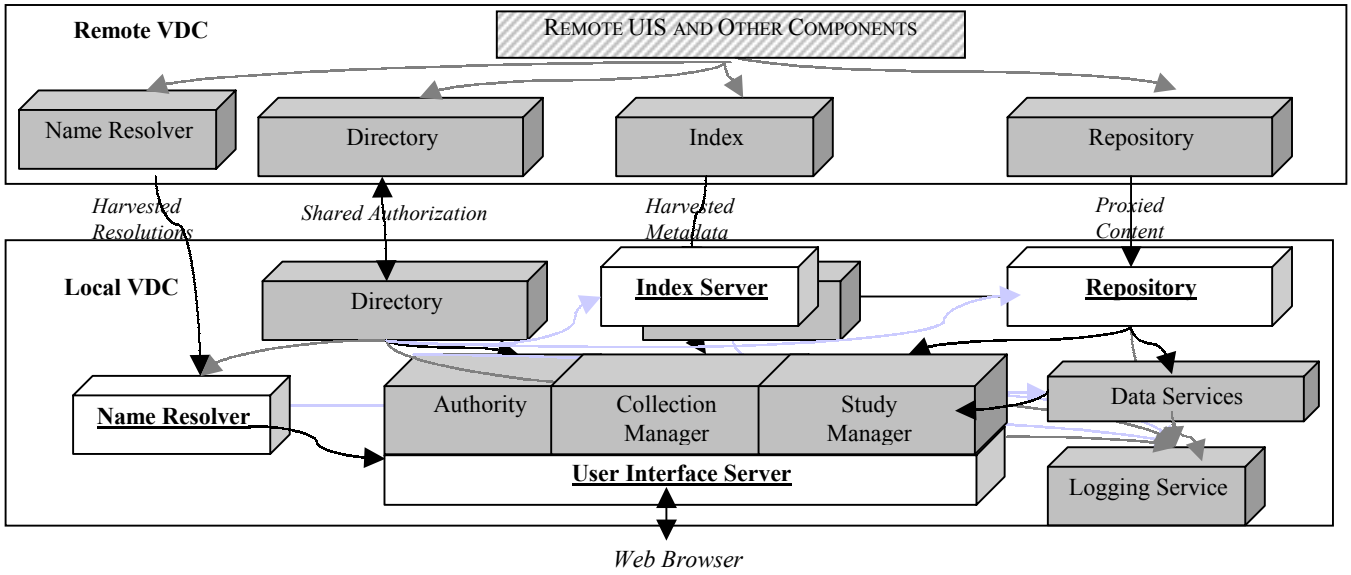

Figure 1. Simplified Representation of VDC

components, we have added to each VDC node a directory server, which allows the components to locate each other, and a centralized logging service, which aids the administrator in tracking usage of the system. Second, the idea of the collection has been expanded to provide independent encapsulation of the specification of virtual content and the 'view', or organization, of that content. Third, the metadata harvester, caching repository proxy, and distributed authentication and authorization components work together support the 'federation' of independent VDC libraries: The result is that content from a group of libraries is made available to their collective users, while each library maintains complete control over how its collections are accessed and how its patrons are authenticated.

Our implementation strategy emphasizes 'open source' development, and integration of the system into a production environment. The director of the *Digital Library Initiative, Phase 2*, of which the VDC is a part, notes the 'unnatural separation' between the producers and consumers of digital libraries, and calls for a balance among research, application, content and collections. [Griffin 1998] In keeping with this admonition, the VDC software system is not simply an isolated research project, it is also a part of Harvard University's first generation *production* digital library system – the VDC software and HMDC site support real use by Harvard library patrons. And this project benefits from taking part in an unusually large and decentralized library system, from cross-fertilization with Harvard's own digital library efforts (see [Flecker 2000]), and from being heavily used by the Harvard research community. In addition, to support the academic norms of openness and accessibility associated with research data, we are in keeping with Lessig's [1999] assertion that the 'code' supporting the fundamental infrastructure for citations must be open. Our code is 'open source', and freely available for modification and use. (see <http://TheData.org>)

3. CONCLUSIONS.

By providing a portable software product that makes the process of data archiving and dissemination automatic and standardized, the Virtual Data Center will help researchers and data archives meet the challenges of sharing and using quantitative data. Consequently, we believe that quantitative research will become easier to replicate and extend.

4. ACKNOWLEDGMENTS

Supported by NSF Award #IIS-9874747.

5. REFERENCES

[1] Altman, M., et al., 2001, "The Virtual Data Center," Working paper. <URL:http://thedata.org/publications/>

[2] Arms, W. Y., et al, 1997. "An Architecture for Information in Digital Libraries," *D-Lib Mag*.

[3] Davis, J. R. and C. Lagoze, 2000. "NCSTRL: Design and Deployment of a Globally Distributed Digital Library," *JASIS*, 51(3):273-280.

[4] Flecker, D. 2000. "Harvard's Library Digital Initiative" *D-LIB Mag*. 6(11).

[5] Griffin, S. 1998. "NSF/DARPA/NASA Digital Libraries Initiative" *D-Lib Mag*.4(7)

[6] Lessig, Lawrence 1999. *Code, And Other Laws of Cyberspace*. Basic Books, NY.

[7] Lagoze, C., D. Fielding, November 1998. "Defining Collections in Distributed Digital Libraries," *D-Lib Mag*.

[8] Sieber, J. E. (ed.) (1991). *Sharing Social Science Data*. Sage Publications, Inc, CA.